ICIC Express Letters
Part B: Applications
Volume **9**, Number **3**, March **2018**

ICIC International ©2018 ISSN 2185-2766
pp. **233–240**

SPARSE NON-NEGATIVE MATRIX FACTORIZATION FOR INDEPENDENT FEATURE LEARNING

Weichun Huang[1], Weijian Xie[1], Liyan Xiong[2]
Xiaohui Huang[2] and Xiong He[2]

[1]School of Software
[2]School of Information Engineering
East China Jiaotong University
No. 808, Shuanggang East Avenue, Nanchang 330013, P. R. China
{ hwc1968; xlyecjtu; hexiong362528 }@163.com; 1021737533@qq.com; hxh016@hotmail.com

Received August 2017; accepted November 2017

Abstract. *Non-negative Matrix Factorization (NMF) can learn local information hidden in an object, and is widely applied in the fields of data mining and machine learning. However, NMF does not always achieve superb performances because the features learned by using the non-negative constraint are usually non-orthogonal and overlap in semantics. It is still an open research problem to improve the semantic independence of latent features meanwhile maintaining the interpretability of NMF. In this paper, we propose a novel NMF, called Sparse Non-negative Matrix Factorization for Independent Feature Learning (INMF). The $L_{2,1/2}$ sparse constraint is used in NMF, and cosine similarity of latent features is taken into account. Therefore, it can effectively enhance the discrimination and the semantic independence of NMF. Thus, we design an objective function by combining the objective function of traditional NMF and the sparse and consine constraint conditions for our proposed method. Subsequently, the iterative updating rules are conducted by optimizing the new objective function. Experimental results on document clustering indicate that our algorithm surpasses baseline methods in terms of a set of evaluations based on real datasets.*

Keywords: Non-negative matrix factorization, $L_{2,1/2}$ sparse, Independent feature learning, Cosine similarity

1. **Introduction.** The purpose of NMF is to decompose the original high dimensional data matrix into two low dimensional data matrices, and the product of the two low dimensional data matrices approximates the original high dimensional data matrix as much as possible.

We have an original data matrix $X = [x_1, x_2, \ldots, x_N] \in R_+^{M\times N}$. NMF seeks to decompose X into non-negative basis matrix $U = [u_1, u_2, \ldots, u_K] \in R_+^{M\times K}$ and non-negative coefficient matrix $V = [v_1, v_2, \ldots, v_N] \in R_+^{K\times N}$, where K is the number of latent features. This can also be written as the equivalent vector formula $x_j \approx \sum_{i=1}^{K} u_i V_{ij}$. Usually, we have $K \ll \min(M, N)$ for rank reduction and v_j are the weight coefficient of the original data vector x_j on the columns of U. NMF decomposes data matrix into the linear combination of the basic vectors.

Several cost functions have been used in the literature to measure the quality of NMF. The main issue is to find the factor matrices (U, V) that achieve the minimum of the chosen cost function. Many researchers have worked out improved methods to obtain a solution to NMF, which can be incompletely divided into constrained NMF, structured NMF and generalized NMF. Constrained NMF imposed some additional constraints as regularization to construct loss functions [1-4]. While structured NMF modified the standard factorization formulations such as weighed NMF [5,6]. Generalized NMF broke

through the conventional data types or factorization modes in a broad sense [7-9]. However, the inborn correlations between latent features make the former NMF algorithms be short of discrimination and increase the difficulty of minimizing loss function. In addition, the previous NMF algorithms do not take full advantages of the sparsity of matrices and they neglect the useful information of the correlation between different features.

Sparse feature selection is aimed to apply a variety of sparse models to realizing feature selection and achieving the sparse data representation. Much works [10-12] have extended the L_1-norm to the L_p-norm ($0 < p < 1$) for better sparsity. In [13,14], Xu et al. have concluded when p is $1/2$, the L_p-norm, i.e., $L_{1/2}$-norm has the best sparsity. In [15], Nie et al. have introduced a joint $L_{2,1}$-norm minimization on both loss function and regularization for feature selection. However, $L_{2,1}$-norm has not good sparsity because it is based on L_1-norm. Recently, Wang and Chen [16] have proposed an idea to extend $L_{2,1}$-norm to $L_{2,p}$-matrix norm ($0 < p \leq 1$) so as to select joint, more sparse features; at the same time, this model has better robustness than $L_{2,1}$-norm. When p is equal to $1/2$, the $L_{2,p}$-matrix norm has the best performance, so we apply the $L_{2,1/2}$-matrix norm model to our new NMF for sparse constraint.

In this paper, we propose INMF that utilizes cosine similarity to improve the independent feature learning ability of NMF by reducing the correlations between latent features. Further, we introduce $L_{2,1/2}$ sparse constraint into INMF. Thus the semantic information in latent features is more distinct and the representations in latent space are more discriminative. We compare our methods to several baseline NMF models on document clustering and give the experimental results of our algorithm and other related algorithms on real dataset. The main contributions of our work can be summarized as follows.

(1) INMF can improve NMF by preventing feature co-adaption with cosine similarity adopted.
(2) Sparse constraint with $L_{2,1/2}$-norm on the basic matrix in the feature space is incorporated as the additional condition, which can not only achieve the sparse data representation and simplify the calculation, but also enhance the local learning ability and robustness of the algorithm.

The rest of the paper is organized as follows. In Section 2, we discuss related work. Section 3 introduces our algorithm. Experimental results on clustering are presented in Section 4. Finally, Section 5 concludes our work and provides suggestions for future work.

2. **Related Work.** Research efforts have been made to improve NMF from various perspectives, like constrained NMF [1-4], structured NMF [5,6] and generalized NMF [7-9]. The most common constrained NMF is sparse NMF that the sparseness constraint is helpful in improving the uniqueness of the decomposition along with enforcing a local-based representation, which is typically measured via L_1-norm [1]. Orthogonal NMFs achieve good performances because the result of orthogonal NMF corresponds to a unique sparse area in the solution region, which learns the most distinct [2]. Graph regularized NMF (GRNMF) improved performance in tasks like document and image clusterings, and it modeled the manifold structure by constructing a nearest neighborhood graph on a scatter of data points [3,4]. Weighed formulations are commonly modified versions of learning algorithms, which can be utilized to emphasize the relative importance of different components. Weighted NMFs are popular in collaborative filtering and clustering tasks as they incorporate prior knowledge into loss function according to connections of instances [5,6]. Generalized NMF like semi-NMF [7], non-negative tensor factorization [8], and non-negative matrix-set factorization [9] are proposed for tasks with complicated and heterogeneous information sources.

A special method dropout NMF prevents the co-adaption of hidden units by changing the update process of latent features [17]. Since the stationary co-occurrence is broken, hidden units can still learn from others but with less dependence. Inspired by these, we

propose a new NMF that works by minimizing cosine similarity between latent features. In the following section, we find that NMF can be improved by breaking the correlations between latent features. Hence we incorporate cosine similarity into NMF. Moreover, $L_{2,1/2}$ sparse constraint on the factor U is used to select the most discriminative sparse features.

3. **Methodology.** The purpose of NMF is to let the product of the coefficient matrix U and the basic matrix V approximate the original data matrix X as far as possible. Formula of NMF is as follows:

$$X \approx UV^T \tag{1}$$

We should minimize the squared Euclidean distance loss function:

$$L = \left\|X - UV^T\right\|_F^2, \quad \text{s.t. } U \in R_+^{M\times K},\ V \in R_+^{K\times N} \tag{2}$$

In this paper, we adopt the squared Euclidean distance loss function:

$$L = \left\|X - UV^T\right\|_2^2 = Tr\left(X^TX\right) - 2Tr\left(X^TUV\right) + Tr\left(V^TU^TUV\right) \tag{3}$$

$\|\cdot\|_2^2$ is the squared L_2-norm and $Tr(\cdot)$ represents the trace of matrix. The iterative updating rules of gradient descent algorithm are shown as follows:

$$u_{mk} \leftarrow u_{mk}\frac{\left(XV^T\right)_{mk}}{\left(UVV^T\right)_{mk}}, \quad v_{kn} \leftarrow v_{kn}\frac{\left(U^TX\right)_{kn}}{\left(U^TUV\right)_{kn}} \tag{4}$$

3.1. **Sparse non-negative matrix factorization for independent feature learning.** NMF can be formulated as a linear neural network as the input x_i is represented by a linear combination of base vectors in U:

$$x_i = Uv_i = \sum_k v_{ki}u_k \tag{5}$$

u_k is the k-th latent feature. The innate correlation between latent features influences the optimization process and increases the difficulty of minimizing loss function; unfortunately, it cannot be avoided.

Since latent features are correlated in NMF, co-adaption is a state that the updates of $\{u_k\}_{k=1}^K$ stop at a saddle point, where L can still be further optimized until it reaches the max iteration. Researchers tried to avoid co-adaption by conducting many separate NMFs on the same dataset with different initialization strategies, which bring more computationally expensive [18]. Therefore, we propose a new NMF that in each iteration latent features will gradually be uncorrelated directly by minimizing cosine similarity between latent features. In such a case, u_k will be independently updated and the influence of correlations will be reduced to the least extent.

3.1.1. *Sparse constraint.* Although NMF can reduce the dimensionality of the original data, how to select discriminative features and achieve the sparse representation of data is still complex. Therefore, sparse constraint with $L_{2,1/2}$-norm on the basic matrix U is incorporated as the additional condition. It can be written as:

$$\|U\|_{2,1/2} = \left(\sum_{i=1}^m \|U_i\|_2^{1/2}\right)^2 \tag{6}$$

3.1.2. *Cosine similarity measure.* Theodoridis and Koutroumbas [19] defined cosine similarity measure as $S_{\text{cosine}}(x,y) = \frac{x^Ty}{\|x\|\|y\|}$ where $\|x\| = \sqrt{\sum_{i=1}^l x_i^2}$ and $\|y\| = \sqrt{\sum_{i=1}^l y_i^2}$ are the lengths of the vectors x and y, respectively. Both x and y are l-dimensional vectors. Since cosine measure is easy to interpret and simple to compute for sparse vectors, it is widely used in text mining and information retrieval [20].

3.1.3. *Iterative updating rules.* Taking the above factors into account, we define the objective function of sparse non-negative matrix factorization for independent feature learning as follows:

$$L = \|X - UV\|_F^2 + \theta\|U\|_{2,1/2}^{1/2} + \alpha\|V\|_2 + \beta\sum_{i,j}\cos(U_i, U_j) \tag{7}$$

θ, α, β are non-negative which can balance the several terms and the weight of the last reconstruction error term. θ, α are the sparse parameters and β is the cosine similarity parameter.

In order to optimize this objective function, we can translate the objective function in Formula (7) as follows:

$$\begin{aligned} L = {} & Tr\left(X^TX\right) - 2Tr\left(X^TUV\right) + Tr\left(V^TU^TUV\right) + \alpha\left(Tr\left(V^TV\right)\right) \\ & + 4\theta Tr\left(U^TQU\right) + \beta Tr\left(U^TUS\right) \end{aligned} \tag{8}$$

$Q = [q_{ij}] \in R^{m\times m}$ is a diagonal matrix. We can calculate the i-th diagonal element q_{ij} of the diagonal matrix Q as follows:

$$q_{ij} = \frac{1}{4\|U_i\|_2^{3/2}} \tag{9}$$

In order to avoid overflow, we add a small enough constant ε into the definition of the matrix Q, so Formula (8) can be rewritten as follows:

$$q_{ij} = \frac{1}{4\max\left(\|U_i\|_2^{3/2}, \varepsilon\right)} \tag{10}$$

where $S = [s_{ij}] \in R^{k\times k}$, we can calculate the elements s_{ij} of the matrix S as follows:

$$s_{ij} = \sum_{i,j}^{k}\cos\left(U_i, U_j\right) \tag{11}$$

In order to obtain the iterative updating rules of the basic matrix U and the coefficient matrix V, we should take the partial derivatives of L:

$$\frac{\partial L}{\partial U} = -2XV^T + 2UVV^T + 8\theta QU + 2\beta US \tag{12}$$

$$\frac{\partial L}{\partial V} = -2U^TX + 2U^TUV + 2\alpha V \tag{13}$$

The iterative updating rules of the basic matrix U and the coefficient matrix V are shown as follows:

$$U_{mk} \leftarrow U_{mk}\frac{\left(XV^T\right)_{mk}}{\left(UVV^T + 4\theta QU + \beta US\right)_{mk}} \tag{14}$$

$$V_{kn} \leftarrow V_{kn}\frac{\left(U^TX\right)_{kn}}{\left(U^TUV + \alpha V\right)_{kn}} \tag{15}$$

3.2. **Convergence analysis.** In this section, we analyze the convergence of our algorithm and prove the objective function in Formula (7) decreases monotonically in the iterative updating rules (14) and (15).

We analyze the convergence of the iterative updating rule (14).

Lemma 3.1. [21]

$$\sum_{i=1}^{m}\left(\left\|g_i^{t+1}\right\|_2^{1/2} - \frac{\left\|g_i^{t+1}\right\|_2^2}{4\left\|g_i^t\right\|_2^{3/2}}\right) \le \sum_{i=1}^{m}\left(\left\|g_i^t\right\|_2^{1/2} - \frac{\left\|g_i^t\right\|_2^2}{4\left\|g_i^t\right\|_2^{3/2}}\right) \tag{16}$$

Proof: From Lemma 3.1, we have:

$$\sum_{i=1}^{m}\left(\left\|U_i^{t+1}\right\|_2^{1/2}-\frac{\left\|U_i^{t+1}\right\|_2^2}{4\left\|U_i^t\right\|_2^{3/2}}\right)\leq\sum_{i=1}^{m}\left(\left\|U_i^t\right\|_2^{1/2}-\frac{\left\|U_i^t\right\|_2^2}{4\left\|U_i^t\right\|_2^{3/2}}\right) \tag{17}$$

We define a function as follows:

$$\begin{aligned}H(U,V)=&Tr\left(X^TX\right)-2Tr\left(X^TUV\right)+Tr\left(V^TU^TUV\right)\\&+\alpha\left(Tr\left(V^TV\right)\right)+\beta Tr\left(U^TUS\right)\end{aligned} \tag{18}$$

Since $\|U\|_{2,1/2}^{1/2}=\sum_{i=1}^{m}\|U_i\|_2^{1/2}$, we can obtain the following inequation:

$$\begin{aligned}&H^{t+1}+\theta\sum_{i=1}^{m}\frac{\left\|U_i^{t+1}\right\|_2^2}{4\left\|U_i^t\right\|_2^{3/2}}=H^{t+1}+\theta\left\|U^{t+1}\right\|_{2,1/2}^{1/2}+\theta\sum_{i=1}^{m}\left(\frac{\left\|U_i^{t+1}\right\|_2^2}{4\left\|U_i^t\right\|_2^{3/2}}-\left\|U_i^{t+1}\right\|_2^{1/2}\right)\\&\leq H^t+\theta\sum_{i=1}^{m}\frac{\left\|U_i^t\right\|_2^2}{4\left\|U_i^t\right\|_2^{3/2}}=H^t+\theta\left\|U^t\right\|_{2,1/2}^{1/2}+\theta\sum_{i=1}^{m}\left(\frac{\left\|U_i^t\right\|_2^2}{4\left\|U_i^t\right\|_2^{3/2}}-\left\|U_i^t\right\|_2^{1/2}\right)\end{aligned} \tag{19}$$

Combining (15) with (18), we can get the following inequation:

$$H^{t+1}+\theta\left\|U^{t+1}\right\|_{2,1/2}^{1/2}\leq H^t+\theta\left\|U^t\right\|_{2,1/2}^{1/2} \tag{20}$$

Therefore, the objective function in Formula (6) decreases monotonically in the iterative updating rule (14).

Nextly, we analyze the convergence of the iterative updating rule (15).

Definition 3.1. *$G(h,h')$ is an auxiliary function for $F(h)$ if the following conditions are satisfied.*

$$G(h,h')\geq F(h),\quad G(h,h)=F(h) \tag{21}$$

Lemma 3.2. *If G is an auxiliary function, then F is nonincreasing under the update*

$$h^{t+1}=\underset{h}{\operatorname{argmin}}G\left(h,h^t\right) \tag{22}$$

Obviously, it can prove $F\left(h^{t+1}\right)\leq G\left(h^{t+1},h^t\right)\leq G\left(h^t,h^t\right)=F\left(h^t\right)$ and $F(h)$ is convergent.

Lemma 3.3.

$$G\left(V,V^t\right)=F\left(V^t\right)+F'\left(V^t\right)\left(V-V^t\right)+\frac{U^TUV+\alpha V}{V^t}\left(V-V^t\right)^2 \tag{23}$$

is the auxiliary function for $F(V)$.

Proof: The first-order derivative and second-order derivative on $F(V)$ are $F'(V)=\left(-2U^TX+2U^TUV+2\alpha V\right)_{kj}$, $F''(V)=\left(2U^TU+2\alpha\right)_{kk}$, so the Taylor expansion of $F(V)$ can be measured as follows:

$$F(V)=F\left(V^t\right)+F'\left(V^t\right)\left(V-V^t\right)+\left(U^TU+\alpha\right)\left(V-V^t\right)^2 \tag{24}$$

Since

$$\left(U^TUV\right)_{kj}=\sum_h\left(U^TU\right)_{kh}V_{hj}^t\geq\left(U^TU\right)_{kk}V_{kj}^t$$

$$(\alpha V)_{kj}=\alpha V_{kj}^t$$

we have

$$\frac{U^TUV+\alpha V}{V^t}\geq U^TU+\alpha$$

so that $G\left(V,V^t\right)\geq F(V)$.

According to the simultaneous Equations (22) and (23), we know $G\left(V^{t+1},V^t\right)$ is the local minimum of (23) and V^{t+1} is the corresponding local minimum point.

$G(V, V^t)$ is the auxiliary function for $F(V)$, so F decreases monotonically by Formula (15).

4. Experimental Results.

4.1. Datasets.

The corpus is fetch_20newsgroups, which is a collection of about 20,000 news documents, partitioned into 20 different groups. It was originally collected by Lang Ken and contains 18,846 documents and only 1,000 words after we preprocess were reserved in this paper. Each document is converted to a vector $x_t \in R^{1000}$, leading to the data matrix $X \in R^{18846\times1000}$. We apply the standard NMF, sparse NMF (SNMF), and our new algorithm to this X, for comparative study in terms of the Precision and the Recall.

4.2. Experimental settings.

Evaluation Metrics. In this paper, we use three evaluation metrics, i.e., Precision, Recall, and F1-score to estimate the clustering performance of the above clustering algorithms.

The formula is as follows:

$$\text{Precision} = \frac{n_{ij}}{n_j} \qquad \text{Recall} = \frac{n_{ij}}{n_{j'}} \tag{25}$$

Among them, n_{ij} is the number that documents in a known class i belong to the cluster j, n_j is the number of documents in the known class i, n'_j is the number of documents in the cluster j.

The F1-score considers both the precision p and the recall r of the documents to compute the score:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

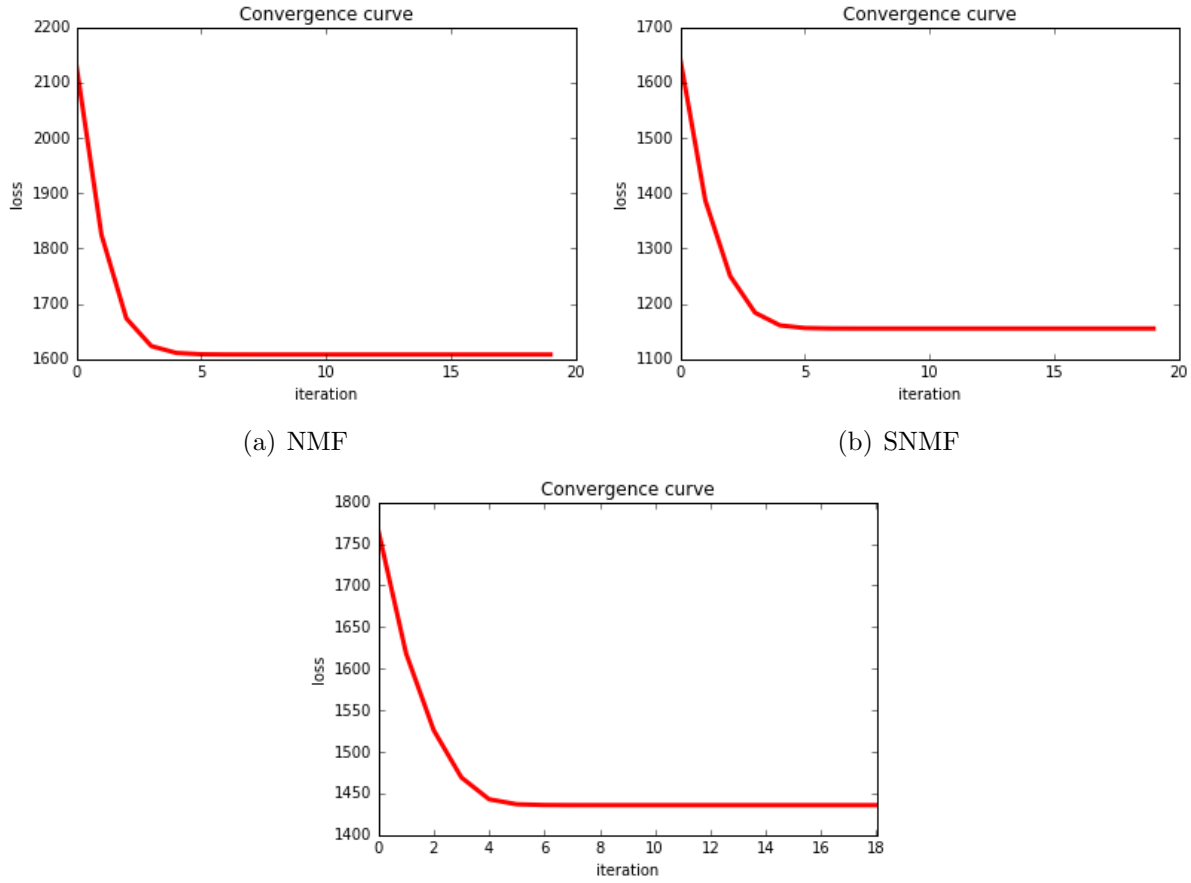
Parameter Settings. As to fetch_20newsgroups dataset, the parameters α, β, θ for new NMF are selected from region $\{0.01 \sim 0.1\}$, and the balance parameters α, β of regularization terms for SNMF are the same as above. To verify the performances on different data sizes, K is set to $\{5, 10, 15, 20\}$ respectively.

4.3. Clustering results.

Table 1 shows the comparisons of two baseline methods with our method on fetch_20newsgroups evaluated by Precision, Recall, and F1-score respectively. The best performances in three versions of methods are boldfaced. It shows that the new method performs better than conventional methods regardless of K although the preponderance is not obvious. It demonstrates the effectiveness of preventing the co-adaption of latent features. Besides, the overall performances decrease with K because a larger dataset with more topics is more difficult for clustering.

TABLE 1. Clustering performances on fetch_20newsgroups

K	Precision			Recall			F1-score		
	NMF	SNMF	INMF	NMF	SNMF	INMF	NMF	SNMF	INMF
5	0.199	0.205	**0.223**	0.198	0.206	**0.222**	0.198	0.206	**0.222**
10	0.100	0.102	**0.109**	0.099	0.101	**0.108**	0.099	0.101	**0.108**
15	0.064	0.066	**0.089**	0.064	0.066	**0.088**	0.064	0.066	**0.088**
20	0.051	0.050	**0.059**	0.051	0.049	**0.058**	0.051	0.050	**0.058**



(a) NMF (b) SNMF

(c) INMF

FIGURE 1. Convergence curves of NMF, SNMF and INMF on fetch_20newsgroups

4.4. **Parameter selection and convergence analysis.** We compare the convergence curves of conventional NMF, SNMF as well as our algorithms on the dataset with $K = 5$, when p is is set to $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ respectively. In Figure 1, the number of iterations is used as X-axis and the objective function value is used as Y-axis. Considering both clustering performance and time consumption, we set the most appropriate parameter as mentioned above. We only show the result when p is a value for each algorithm. All methods tend to converge within 20 iterations. On the fetch_20newsgroups, the numbers of iterations of NMF and SNMF are about 18 and 19 respectively, and the objective function tends to be stable. However, the INMF can converge at 15 iterations. This is because more latent feature is uncorrelated in each iteration, and the accumulated effect leads to a better performance.

5. **Conclusion.** In this paper, we analaze how the correlations among latent features in NMF affect performance and propose INMF called sparse non-negative matrix factorization for independent feature learning which can not only utilize cosine similarity of vectors, but also effectively learn local information of the objectives. In INMF, latent features are updated by minimizing cosine similarity of it in each iteration. Co-adaption is effectively prevented in the proposed algorithm, so latent features are more definite and discriminative. In addition, $L_{2,1/2}$ sparse constraint is incorporated as the additional conditions in the NMF, which can make the basic matrix with a good sparsity. From all the experimental results above, we can make a conclusion that our algorithm is with encouraging performance on both Precision and Recall. In the future, we will explore the INMF with other loss functions and variations. The new method will be put forward to deal

with other applications in different domains, such as video processing, natural language processing.

Acknowledgement. This research was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61363072 and No. 61562027, Science and Technology Department of Jiangxi Province under Grant No. 20161BBI90032, 20151BB990041, 20161BAB212050, Education Department of Jiangxi Province under Grant No. GJJ160508, Graduate Student Innovation Foundation of Jiangxi Province under Grant No. YC2016-S261. The authors also gratefully acknowledge the helpful comments and suggestions provided by the reviewers.

REFERENCES

[1] N. Mohammadiha and A. Leijon, Nonnegative matrix factorization using projected gradient algorithms with sparseness constraints, *Proc. of IEEE International Symposium on Signal Processing and Information Technology*, pp.418-423, 2009.

[2] C. Ding, T. Li, W. Peng and H. Park, Orthogonal nonnegative matrix tri-factorizations for clustering, *Proc. of the 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp.126-135, 2006.

[3] D. Cai, X. He, J. Han and T. Huang, Graph regularized non-negative matrix factorization for data representation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.33, no.8, pp.1548-1560, 2011.

[4] R. Zhi, M. Flierl, Q. Ruan and W. Kleijn, Facial expression recognition based on graph-preserving sparse non-negative matrix factorization, *Proc. of IEEE the 16th Int'l Conf. Image Processing (ICIP)*, pp.3293-3296, 2009.

[5] Y. Mao and L. Saul, Modeling distances in large-scale networks by matrix factorization, *Proc. of ACM SIGCOMM Conf.*, pp.278-287, 2004.

[6] Y. Kim and S. Choi, Weighted nonnegative matrix factorization, *Proc. of IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp.1541-1544, 2009.

[7] C. Ding, T. Li and M. Jordan, Convex and semi-nonnegative matrix factorizations, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.32, no.1, pp.45-55, 2010.

[8] M. Mørup, L. Hansen and S. Arnfred, Algorithms for sparse nonnegative tucker decompositions, *Neural Computation*, vol.20, no.8, pp.2112-2131, 2008.

[9] L. Li and Y.-J. Zhang, Non-negative matrix-set factorization, *Chinese J. Electronics and Information Technology*, vol.31, no.2, pp.255-260, 2009.

[10] S. Foucart and M. J. Lai, The sparest solutions of underdetermined linear system by LQ-minimization for, *Appl. Comput. Harmonic Anal.*, vol.26, no.3, pp.395-407, 2008.

[11] R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, *IEEE Signal Process. Lett.*, vol.14, no.10, pp.707-710, 2007.

[12] R. Chartrand, Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data, *Proc. of IEEE International Symposium on Biomedical Imaging*, pp.262-265, 2009.

[13] Z. B. Xu, X. Y. Chang, F. M. Xu and H. Zhang, $L_{1/2}$ regularization: A thresholding representation theory and a fast solver, *IEEE Trans. Neural Netw. Learn. Syst.*, vol.23, no.7, pp.1013-1027, 2012.

[14] Z. B. Xu, H. Zhang, Y. Wang, X. Y. Chang and Y. Liang, $L_{1/2}$ regularizer, *Sci. China*, vol.53, no.6, pp.1159-1169, 2010.

[15] F. P. Nie, H. Huang, X. Cai and C. Ding, Efficient and robust feature selection via joint $L_{2,1}$-norms minimization, *Proc. of NIPS*, pp.1813-1821, 2010.

[16] L. P. Wang and S. C. Chen, $l_{2,p}$ – Matrix norm and its application in feature selection, *CoRR*, vol.abs/1303.3987, http://arxiv.org/abs/1303.3987, 2013.

[17] Z. He, J. Liu, C. Liu, Y. Wang and A. Yin, *Dropout Non-negative Matrix Factorization for Independent Feature Learning*, Springer International Publishing, 2016.

[18] A. N. Langville, C. D. Meyer, R. Albright, J. Cox and D. Duling, Algorithms, initializations, and convergence for the nonnegative matrix factorization, *Clinical Orthopaedics and Related Research*, 2014.

[19] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 2008.

[20] I. S. Dhillon and D. S. Modha, Concept decompositions for large sparse text data using clustering, *Math. Learn.*, vol.42, nos.1-2, pp.143-175, 2001.

[21] C. Shi, Q. Ruan, G. An et al., Hessian semi-supervised sparse feature selection based on $L_{2,1/2}$-matrix norm, *IEEE Trans. Multimedia*, vol.17, no.1, pp.16-28, 2015.